## 2. 1. LA VARIABLE '*DISCOURS*'

### 2. 1. 1. Textes et contextes

Dans les débuts des traitements informatiques et statistiques de textes, on a cru pouvoir extrapoler, à partir de recherches sur des corpus de l'ordre de un million de signes, des probabilités d'apparition des mots du "lexique-type de la langue"[1] : on espérait pouvoir affecter à ceux-ci une sorte d'indice statistique, propre à cet élément du lexique, qui viendrait s'ajouter à sa transcription phonologique et à ses caractéristiques morphologiques et syntaxiques. On a vu au chapitre 1 que ces recherches se situaient dans un contexte où l'on utilisait l'ordinateur pour dégager "le vocabulaire fondamental" d'une langue à des fins pédagogiques et idéologiques et que ce contexte était marqué aussi par les recherches cybernétiques, le cadre de la théorie mal nommée "de l'information" où les questions du sens et de l'interprétation humaine sont, paradoxalement, évacuées. Nos recherches en sémantique descriptive sont, elles, centrées sur l'interprétation du sens en contexte, et s'appuient sur des concepts linguistiques utilisés dans le cadre d'une théorie sémiotique du texte. Nous allons observer quels éléments d'information apportent des outils informatiques et statistiques sur les emplois en contexte d'un mot banal du lexique du français, en nous mettant dans la situation heuristique de description pas à pas de corpus.

La question qu'on peut éclairer d'un jour nouveau grâce à des corpus électroniques importants est celle du "voisinage" : en effet, si "la langue" est la même quels que soient le type de texte et le contexte, le test probabiliste devrait sélectionner à peu près les mêmes cooccurrents au voisinage d'un mot identique dans des corpus différents, et leurs scores statistiques devraient se ressembler peu ou prou. Puisque l'ensemble Frantext propose des discours et des genres textuels différents[2], on peut observer la question des associations contextuelles dans des textes variés.

### 2. 1. 2. La question de la lemmatisation

La première étape du traitement statistique commence par la reconnaissance des "unités" constituant le corpus de travail et le corpus de référence, et la question de la lemmatisation, qui a suscité, nous l'avons rappelé, des débats passionnés dans la communauté de la statistique lexicale, se pose alors. La lemmatisation est loin d'être une option en "tout ou rien", car elle est, en fait, une tâche triple, dont chaque étape oblige à confronter des critères théoriques à des

[1] "L'analyse des caractères arithmo-sémantiques du lexique d'un texte implique une liste rang-fréquence de tous les mots d'une part du lexique du texte, d'autre part du lexique-type de la langue qui permettraient d'établir un coefficient de corrélation. Mais nous ne connaissons pas le lexique de la langue ; et c'est une notion des plus confuses puisqu'il n'est même pas possible de définir la langue en général (…) Dans la pratique, il n'est cependant pas interdit de poser le problème. En effet, nous possédons des listes de fréquence qui nous donnent une approximation suffisante pour le français. La plus solide est celle de Vander Beke (en note : George Vander Beke, *French Word Book*, New York, The Mac Millan Cy, 1931) que nous avons adoptée comme système de référence. Elle repose sur la compilation de 88 textes de 13000 mots chacun, soit 1 200 000 ou 600 000 mots forts ("mots de signification", *cf.* p. 62). Ces textes, tous en prose, comprennent 33 romans, 13 pièces de théâtre, 14 journaux et revues, 13 œuvres scientifiques et philosophiques, 16 œuvres historiques et critiques. Les textes partent de Balzac, Michelet, Musset jusqu'à Proust, Bergson, Paul Morand. On peut considérer qu'elle est représentative de la prose littéraire de la fin du XIX-début XXè siècle." (Guiraud P. 1954, 63).

[2] Ce sont des niveaux hiérarchiques de structuration des textes, qui seront explicités plus loin.

réalités empiriques incontournables. Ces étapes consistent à délimiter le signe, à décider de rattacher une forme à un lemme (forme canonique), ce qui pose la question des homographies et ne peut se réaliser qu'en attribuant, parallèlement, une étiquette de partie du discours à cette forme : comme d'une part, les débats sur la définition du "mot" ou sur l'extension de certaines catégories grammaticales ne sont pas tranchés dans la communauté linguistique, et que d'autre part nous nous trouvons devant des masses de textes de près de 200 millions d'unités graphiques, si nous travaillons sur toute la banque Frantext, de 40 millions si l'on choisit le corpus Roman, et de 4,6 si l'on se cantonne à la poésie, parente pauvre, on peut prendre la mesure du problème. Il n'est pas question de choisir l'option la plus satisfaisante intellectuellement qui est d'étiqueter "à la main" chacune des unités, ce qui a été fait, pour les trois corpus français annotés bien connus, *le Menelas* (84839 occurrences et 6191 formes, discours médical, rassemblé pour le projet européen de compréhension de comptes rendus d'hospitalisation), *le Mitterand1* constitué des interventions radiotélévisées du premier septennat de F. Mitterand (Institut d'études politiques de Grenoble, 305124 occurrences, 9309 formes) et du corpus nommé *Enfants*, consistant en textes de réponses à une enquête menée par le Centre de Recherches et de Documentations sur la Consommation (15523 occurrences, ponctuation non comprise, 1305 formes)[3]. Les problèmes concernant l'étiquetage des corpus sont bien détaillés dans Habert et *alii*, 1997 (chapitres 1, 2, et 3 pour l'étiquetage sémantique), qui remarquent : "Cette diversité [des étiquetages] tient à l'utilisation envisagée du corpus mais aussi à son mode d'étiquetage (manuel ou automatique) ainsi qu'à l'absence de consensus sur certaines catégories ou sur leur extension. L'expérience prouve qu'un groupe d'annotateurs n'est pas forcément cohérent dans les étiquettes qu'il attribue manuellement à un texte. Il en va de même pour un individu donné au fil du temps"[4].

Rappelons que le but de notre recherche est l'aide à l'interprétation du contenu, dans le cadre d'une théorie de sémantique linguistique qui postule que le signe isolé n'a pas de sens : la question de l'étiquetage sémantique préalable[5] ne se posait donc pas et celle de la catégorie grammaticale ne revêtait pas non plus d'importance primordiale[6]. Nous avons pu élaborer les programmes dont nous avions besoin à partir des ressources disponibles.

Les programmes du SAAS portent de fait la marque de choix issus des précédents travaux du laboratoire, de l'état des données et des outils "maison" de l'époque. Ils s'appuient en particulier sur des travaux informatiques et statistiques qui avaient été menés dans les années 1970 sur le premier corpus saisi en vue de la rédaction du Dictionnaire, dans le laboratoire nancéien alors nommé "Centre de recherche pour un Trésor de la langue française" : ce corpus, qui comporte 70 millions d'occurrences, a donné lieu à la publication du *Dictionnaire des Fréquences* 1969-71, et a servi de base à différents travaux, dont les recherches sur vastes corpus, inaugurées par É. Brunet. Pour la mise au point des programmes du SAAS, on s'est servi des "dictionnaires de formes" élaborés lors de ces recherches et on a suivi, d'autre part, le choix de ne pas rassembler les formes des participes passé et présent des verbes sous le lemme, à la différence des formes conjuguées (*cf*. dans le tableau ci-dessous *prosterné*, *prostré*,

[3] V. Habert B. et *alii*, 1997, p. 18.
[4] *Loc. cit*. p. 22.
[5] Ce type d'outils porte la marque de la linguistique du signe dont il est issu, *cf*. plus loin.
[6] La sémantique ne s'affranchit pas de la grammaire, ni de la syntaxe (qui n'en est qu'une partie) : nous allons expliciter ci-dessous et au chapitre 3 la méthode d'analyse du sens en termes de traits sémantiques (qui incluent des traits d'accord, de rection, de genre et de nombre, aspectuels, etc.).

*bottant*). Comme on ne disposait pas d'un catégoriseur[7] à l'époque où nous avons effectué ces expériences, on a choisi de traiter les formes homographes[8] comme des lemmes pour ne pas avoir à trancher arbitrairement entre les catégories possibles.

Utiliser un dictionnaire de formes issu d'un corpus antérieur peut avoir introduit certains biais : par exemple, la forme *conculcatrice* (*cf.* 2. 3. 1.) est isolée dans le corpus initial. Si dans l'ensemble textuel Frantext, riche à présent de près de 200 millions d'occurrences, *conculcatrice* reste unique, en revanche *conculcateur* est attesté 4 fois au masculin singulier, dans *Le Roman de la Momie* de Th. Gautier et une fois au pluriel par L. Bloy (l'auteur qui emploie également *conculcatrice* mais dans un autre texte). Le dictionnaire des formes devrait donc rassembler *conculcateur, conculcateurs* et *conculcatrice* sous le lemme *conculcateur*[9], et, dans ce cas, le score statistique de cette forme serait moins élevé, même si, nous l'observerons dans les tableaux suivants, le score constamment important de *conculcatrice* rend compte à la fois de son lien sémantique fort avec le mot pôle, de son caractère de hapax dans le corpus, en même temps que du fait qu'il n'a pas été rattaché au lemme *conculcateur*. Mais ce problème d'un dictionnaire de référence suffisamment adapté au corpus se pose en fait constamment, puisque, comme nous le verrons, la probabilité d'apparition de telle forme du lemme ou de telle forme figée d'un patron varie fortement selon les textes, et que l'on ne peut envisager pour des corpus de cette ampleur une étude exhaustive des formes, en préalable à tout traitement automatique. Selon les degrés d'expérience et les objectifs, les outils disponibles et leur fiabilité, l'étape de lemmatisation d'un corpus peut s'avérer diverse, l'essentiel étant d'adapter les options au type de recherche[10].

### 2. 1. 3. *Pied* et son voisinage dans des corpus contrastés

Dans cette expérience, on a joué sur la variable 'discours' ('littéraire' *vs* 'non littéraire'), en appliquant le test de l'écart réduit à deux corpus, autour du lemme *pied(s)*, qui comporte plus de 16000 attestations uniquement dans le corpus Roman. Comme notre visée était sémantique, nous avons réduit quelque peu le corpus d'observation en ne traitant *pied* que dans les séquences *pied(s) de (d', du*), ce qui représente cependant une masse de contextes importante, puisque, sur les 16186 attestations de *pied* dans le roman (8704 au pluriel et 7482 au singulier), on rencontre 2898 exemples de la séquence *pied(s) de (d', du)*. Ainsi, le corpus de travail pour *pied de + pieds de (du, d')* dans le roman est de 48003 occurrences[11] et le corpus de travail pour *pied de + pieds de (du, d')* dans le technique de 9080 occurrences.

---

[7] Actuellement une partie de Frantext est accessible avec reconnaissance des principales catégories grammaticales (catégorisation effectuée par un outil élaboré par J. Maucourt et M. Papin).

[8] Comme *chaussons*, qui peut être substantif pluriel ou forme verbale, *cf.* tableau 4 du chapitre 3 ; d'après le *Dictionnaire des fréquences* -qui n'a pas été réactualisé sur le corpus total de Frantext- cette forme a une probabilité de 99% d'être substantif.

[9] Dans le TLF, ces mots sont traités sous le mot-vedette *conculcateur* dont l'origine est le latin *conculcare* "fouler avec les pieds, écraser et au fig., fouler aux pieds, opprimer, mépriser" ; *cf.* sur l'axe sémantique "prééminence de X" l'attestation p. 327 du *Roman de la Momie*, l'adresse de Tahoser à Pharaon : - *Oui, tu es le conculcateur des peuples, le dominateur des trônes, et les hommes sont devant toi comme les grains de sable que soulève le vent du Sud.*

[10] Sur la différence entre les choix empiriques liés aux applications et les options uniquement descriptives ou théoriques, v. Fradin 1993 : on peut s'étonner que trop peu de recherches théoriques profitent de l'existence des corpus importants qui apportent pourtant des faits observables et quantifiables.

[11] Une précision quant à ces chiffres : comme le point arrête la sélection automatique, tous les énoncés autour d'une forme du lemme *pied(s)* ne comportent pas 21 occurrences, la proportion n'est que de 83 % de la fenêtre attendue.

Le corpus de référence "Roman" est riche de près de 40 millions de signes, et le corpus "Technique" compte 19 millions de signes. La fenêtre de sélection est constituée, comme nous l'avons signalé dans le chapitre 1 (*cf.* 1. 3) de 10 mots avant et 10 mots après le mot-pivot.

Nous avons signalé au chapitre 1 que Frantext est un ensemble textuel important, à dominante littéraire et dont de nombreux textes littéraires sont donnés dans une bonne édition, au plan philologique, mais que ce corpus est pratiquement brut, car on y accède par les seules informations textuelles suivantes : auteur, titre, forme, séquence de formes, combinaison de formes/séquences avec options de position, formes conjuguées d'un verbe, forme d'un lemme, liste de mots[12]. Le texte est pris comme un tout, (même si des jalons de pages existent) c'est-à-dire qu'il n'y a pas d'accès aux pages ni aux parties du texte, comme les chapitres[13], et encore moins à des divisions internes comme le paragraphe. A l'heure actuelle, cet état semble bien insatisfaisant au regard des besoins nouveaux que les utilisateurs peuvent ressentir devant des masses textuelles qualifiées de "facilement accessibles", mais c'est une contrainte de fait sur tous les ensembles importants et nous aurons l'occasion de mettre en évidence les limites dues à ces lacunes, aussi bien que la richesse des trouvailles que l'on peut faire dès qu'on a affaire à des corpus "homogènes", c'est-à-dire denses au plan sémantique.

### *Le lemme* **pied** *et ses cooccurrents dans le corpus Roman et dans le corpus Technique de Frantext*

Le tableau 1 donne les résultats de la sélection autour du lemme, (dans la séquence *pied(s) de (d', du)* pour les 20 premiers cooccurrents, que nous allons étudier en détail[14] ; on trouvera en annexe un tableau regroupant les 50 premiers cooccurrents de chaque corpus, avec mention du score dans l'autre corpus, s'il y a lieu : on remarque que sur les 50 premiers cooccurrents du corpus Roman, 16 seulement sont également sélectionnés dans le Technique et que sur les 50 premiers cooccurrents du corpus Technique seuls 13 sont également sélectionnés dans le Roman[15]. Pour le corpus de travail Roman, le test a sélectionné 639 cooccurrents et pour le corpus Technique 212 cooccurrents.

---

[12] Il peut y avoir confusion pour des homonymes, à cause de l'état de la base bibliographique : par exemple, le système sélectionne ensemble Guez de Balzac et H. de Balzac à la requête "Balzac", amalgame que l'on peut éviter par l'ajout d'un critère de date ; c'est le cas également de "Duras", qui sélectionne Marguerite Duras avec Mme de Duras (1778-1828). Pour les titres, ce cas peut se produire : le système bibliographique ne souligne pas que la banque possède deux versions de *La Tentation de Saint Antoine* de Flaubert (1849 et 1856), ce qui peut être gênant pour certaines recherches. Nous verrons au chapitre 5 comment le conjugateur intégré peut occasionner du bruit.

[13] Lors des recherches, on voit apparaître parfois le titre du chapitre, mais il peut être fautif, car ces subdivisions ne sont pas gérées par un système cohérent.

[14] La présentation des énoncés est quelque peu différente de celle du chapitre 3 parce que les traitements ont été faits à des époques diverses de la conception des programmes, que nous avons mis au point, testés et améliorés au fur et à mesure des expériences. En effet, les programmes ont été constamment testés en "grandeur réelle", dans le dialogue informaticien-linguiste, et améliorés pour mieux remplir leur tâche d'aide à l'interprétation. C'est ainsi que nous ne disposions pas, dans la première étape, de la formule "sorties des résultats sémantiquement denses, en format paysage, avec mise en majuscules des cooccurrents sélectionnés", mais d'une formule antérieure, réitérant les contextes autant de fois qu'il y a de cooccurrents. Cette formule a été abandonnée parce qu'elle était beaucoup plus longue à dépouiller, d'une part, et faisait moins bien ressortir les traits sémantiques importants dans les associations. La version des contextes en format paysage oblige à avoir des références abrégées pour les textes (en 6 caractères au maximum, *cf.* annexes du chapitre 3), travail qui a été fait pour le corpus roman sur lequel nous avons choisi de mener les différents types de recherches, mais pas pour le corpus technique, qui s'est avéré trop hétérogène et n'a servi de corpus de contraste que dans cette expérience. Cependant nous avons mis en caractères gras les autres cooccurrents sélectionnés dans les énoncés de 10 mots avant et après le pôle.

[15] Un certain nombre de travaux soulignent les différences dans la probabilité qu'a un mot X d'apparaître dans un corpus : Garrigues M., 1998, p. 328 fait observer que, dans un important corpus de textes journalistiques récents (Le Monde, 1992-1993, près de 45 millions de mots), les mots suivants n'apparaissent qu'une fois : *abaque, aboutissons, abondement, abattraient, abordions.* Dans un corpus de romans de 1800 à 1998 de Frantext, qui représente plus de 67 millions de signes, on trouve ces mots avec la fréquence suivante : *abaque* 1, *aboutissons* 2,

Dans ce chapitre, nous allons montrer dans quelles sortes de contextes se situent les associations entre *pied* et le cooccurrent sélectionné, en nous centrant sur les ressemblances et les différences, au plan sémantique, entre les deux corpus, littéraire et non littéraire. A partir du chapitre 3, nous ne travaillons plus que sur le roman et les résultats du corpus roman, autour de *pied(s)* seront examinés plus à fond à cet endroit, où sont présentés les variations entre lemme et forme, et celles suscitées par le changement de la fenêtre de sélection.

| **CORPUS ROMAN** | | **CORPUS TECHNIQUE** | |
|---|---|---|---|
| score | lemme | score | lemme |
| 84 | grue | 76 | bielle |
| 67 | lit | 56 | égalité |
| 51 | conculcatrice | 52 | bottant |
| 36 | astérion | 45 | géranium |
| 36 | mur | 43 | biche |
| 34 | arbre | 37 | coussinet |
| 29 | égalité | 27 | mortaise |
| 26 | diamètre | 26 | circonvolution |
| 24 | escalier | 26 | boeuf |
| 24 | mont | 26 | prostré |
| 24 | prosterné | 25 | bastion |
| 23 | montagne | 24 | équidistant |
| 23 | biche | 23 | falaise |
| 23 | muraille | 23 | pied |
| 23 | mesurer | 21 | tenon |
| 22 | céleri | 21 | croix |
| 22 | croix | 21 | rameuse |
| 21 | échelle | 19 | réparation |
| 20 | assis | 18 | frontal |
| 20 | rocher | 18 | coup |

*Tableau 1 : scores statistiques des 20 premiers lemmes associés à* pied(s) de *dans le corpus Roman*[16] *et dans le corpus Technique.*

### *Les contextes des cooccurrents différents parmi les 20 premiers*

**Les exemples du corpus Technique pour la séquence *pied(s) de (d', du)***

1) dans la plupart des cas, les **coussinets** de pied de **bielle** sont exécutés en **bronze** Ambroise E. Pour le monteur mécanicien, 1949, p. 31

2) **pistons** et parois de cylindres, -axes de pied de **bielle** dans la **bielle**, **-têtes** de **bielle** Chapelain Ch., Cours de technique automobile, 1956, p. 157

3) vérifier aux deux points morts comment se présente le pied de **bielle** par rapport à la **crosse** de **piston** Ambroise E. Pour le monteur mécanicien, 1949, p. 30

Les cooccurrents n$^{os}$ 1 et 6, *bielle* et *coussinet*, proviennent de deux textes concernant le champ de l'automobile (domaines de la technique de conduite et de la mécanique) qui ont entraîné également la sélection de *piston* (score 13), *crosse* (5) et *bronze* (score 3, attesté seulement dans cet énoncé[17] et dans le syntagme *pied de bronze de la statue*, dans un traité de géologie), *conducteur, accélérateur, bouton* et *manivelle* (dans *bouton de manivelle*[18]), *pédale,*

*abondement* 0, *abattraient* 2, *abordions* 6, tandis que pour le même espace de temps, dans le corpus poésie (4,6 millions de signes) on ne trouve aucune attestation de ces formes sauf *abattraient*, avec un seul exemple.

[16] Nous avons éliminé de ce tableau le cooccurrent de rang 16 dans le roman, *opposite*, dû à un doublon ; il figure dans le tableau 5 du chapitre 3 et dans les tableaux d'annexes.

[17] *Bronze* est également sélectionné dans le corpus Roman, avec le score 11, dans des descriptions d'objets dont le pied est constitué de cette matière.

[18] Ce syntagme, datant de 1949 serait-il attesté aujourd'hui dans un ouvrage du même type ?

etc. Ci-dessous, on voit que *bottant* (n° 3), *pied*[19] (n° 14), *réparation* (n° 18), *coup* (n° 20), sont eux issus d'un texte intitulé *Le Football*, et sur ce "sujet" on trouve les cooccurrents *appui* (score 11), *ballon* (score 13), *but*[20], (score 4. 9), ainsi que *joueur*, *montants*, etc. *Coup* est attesté 21 fois dans le corpus de travail, dont une seulement dans un texte sur la menuiserie dans le syntagme *quelques coups de varlope*, tous les autres exemples venant de ce texte sur le football, dans les syntagmes *coup de pied de but, de coin, de réparation*, etc.

4) dont fait partie le j**oueur bottant** le **coup** de pied de **but** devront se trouver en dehors de la surface Mercier J., Le football, 1966, p. 29

5) dont fait partie le j**oueur bottant** le **coup** de pied de **coin** ne pourront s'approcher à moins de Mercier J., Le football, 1966, p. 30

6) un **coup** de pied de r**éparation** pourra être accordé quelle que soit la **position** Mercier J., Le football, 1966, p. 25

7) de passer entre les **montants** sur un **coup** de pied de **réparation** accordé à l'expiration de la première **mi**-temps Mercier J., Le football, 1966, p. 28

8) devant la **jambe** d'**appui** afin de permettre au pied de contact de contrôler le **ballon** en l'**entraînant** dans Mercier J., Le football, 1966, p. 36

En dehors de *pied*, tous ces cooccurrents sont propres au corpus Technique.

*Jambe* (n° 36), sélectionné dans le texte sur le football (ex. 8), l'est aussi dans un traité sur la danse :

9) le danseur tourne sur lui-même, le pied de position servant de pivot, l'autre **jambe** étant Bourgat M., Technique de la danse, 1959, p. 107

et un texte intitulé *L'Education physique et sportive* ; on le lit également dans *l'Encyclopédie médicale Quillet*, et, dans les *Entretiens sur l'Architecture* de Viollet-Le-Duc, dans le syntagme *jambes de force*, où l'on comprend qu'il ne s'agit pas de l'anatomie humaine. Dans le corpus Roman, il est retenu avec le score 6, et figure dans des descriptions d'attitudes humaines uniquement.

On a vu (exemple 7) que *mi* était sélectionné, en fait pour *mi temps*[21] : ce morphème est attesté également comme substantif, un fois, dans le domaine de la musique :

10) donnant à l'oreille le **mi** du 16è pied de l'orgue ; et à l'**aigu** par le Arts et Littérature, société contemporaine, 1935, p. 4002

Un texte sur la gastronomie française est à l'origine de la sélection de *carotte, bœuf, veau* :

11) douzaine de **carottes** rouges divisées en fragments et deux pieds de **veau**, recouvrez le tout de bouillon de **bœuf** Grandes heures de la cuis fr., 1964, p. 201

et aussi de *bouillon, céleri, oignon*. Mais *bœuf* est attesté de manière inattendue pour le lecteur moyen d'aujourd'hui dans le syntagme *pied de bœuf* qui désigne une sorte de jeu[22]

12) iv. Représentation artistique du jeu du pied de **bœuf**. Allemagne H.-R., Récréations et passe-temps, 1904, p. 205

De ces cooccurrents, seuls *céleri* (*cf.* ci-dessous, ex. b', c', d') et *veau* sont sortis par le test pour le Roman, avec le score 4 pour *veau* qui figure dans les syntagmes *le veau d'or* (*sous les pieds d'Abraham, de Moïse*) et *pied(s) de veau* (chez Zola et Proust).

---

[19] Il s'agit de la sélection de *pied* comme cooccurrent, se trouvant dans la fenêtre de sélection près de *pied* comme mot pôle. *Pied* est sélectionné également dans les domaines de la géométrie (*pied de la perpendiculaire, [pied d'une] oblique*), de la danse (*écart glissé de trois demi-pieds de talon ; ramener les pieds* …), de la menuiserie (*pieds de devant, de derrière, pieds en console*), dans un traité sur l'histoire des jouets et un texte sur l'histoire des techniques et des inventions, pour l'unité de mesure (*une ouverture de 4 pieds de long ; à 100 pieds de distance*), tandis qu'il est question du *pied central* et du *pied d'angle* d'une table de salle à manger dans le *Larousse ménager illustré* de 1926 et qu'on trouve : *le fleuve se resserre au pied du roc* dans un ouvrage de géographie ; un autre texte de sciences humaines, sur le fédéralisme européen emploie *le pied de l'égalité*, locution dont nous reparlons plus loin, et *le pied de la proportionnalité.*

[20] Cependant *buter* sélectionné avec le score important de 12, n'a pas la "couleur" du football : il est attesté dans un texte sur l'architecture faisant renvoi à une illustration (Viollet-Le -Duc, *loc. cit.* : *un petit épaulement d*, pour buter le pied de la pièce butante en fonte g**) et au sens figuré de "se heurter à des difficultés" dans un texte de sciences humaines, *L'Univers économique et social*, 1960.

[21] Nous aurons l'occasion de voir à différentes reprises comment ce problème de la reconnaissance correcte des unités de la langue est important ; si *mi-temps* avait été reconnu comme lexie, il n'aurait pas été sélectionné car son score de cooccurrence avec *pied* n'est que de 1 et le substantif de musique ne l'aurait pas été non plus, pour les mêmes raisons. Ni *mi*, ni *a fortiori mi-temps* ne sont relevés dans le Roman.

[22] Il fait partie des jeux de société dont l'intérêt vient des gages et pénitences donnés au perdant.

Voici les contextes des cooccurrents n$^{os}$ 7, 10 et 11 :

13) contre ce **bastion**, au pied de ce **bastion**, sur ce **bastion** se sont rencontrées Brunhes J., La géographie humaine, 1942, p. 278
14) par l'autre à **tenon** et à **mortaise** dans le pied du bureau Nosban, Manuel menuis. ébénist., marquet., 1857, p. 92
15) aboutir, finalement, au groupe de *marie au pied de la **croix** : tantôt **prostrée**, à moitié évanouie Febvre L., Combats pour l'histoire, 1952, p. 233
16) le thème de la vierge **prostrée** et pleurante au pied de la **croix** dans l'*espagne pathétique de la fin Febvre L., Combats pour l'histoire, 1952, p. 233

Si l'exemple n°14 "sonne" technique, 13, 15 et 16, issus de textes de sciences humaines, se rapprochent plutôt des textes du corpus Roman par le vocabulaire comme *prostrée, pleurante, la Vierge, au pied de la croix* (qui ne sont pas attestés ailleurs dans le corpus technique), et par des aspects stylistiques aussi, comme le rythme ample de 15 et 16, et la répétition en rythme ternaire pour 13. Nous verrons dans la partie ci-dessous les raisons de cette discordance de ton.

Sur les 45 exemples de *circonvolution* (n° 8) dans le corpus de référence, 4 se trouvent en cooccurrence avec *pied*, dans deux textes du domaine médical, et ce, uniquement dans le syntagme *pied de la* [2$^{ième}$ ou 3$^{ième}$] *circonvolution frontale (gauche)*, non attesté dans le corpus Roman :

17) lorsqu'un individu porte une **lésion** localisée au pied de la 3è **circonvolution frontale gauche**, il Camefort-Gama, Sciences naturelles, 1960, p. 268
18) images du langage, localisation du langage articulé au pied de la 3è **circonvolution frontale**, Ce que la France a apporté à la méd., 1943, p. 251

On a vu que *pied de céleri* était attesté dans le domaine de la cuisine ; quant au syntagme *pied de géranium*, très proche au plan linguistique, on a la surprise de le lire dans un ouvrage médical, trois fois, et uniquement là :

19) un pied de **géranium** est exposé à la lumière pendant plusieurs heures Camefort-Gama, Sciences naturelles, 1960, p. 323

Qu'en est-il de mots que l'on ne qualifierait pas *a priori* de "termes techniques" ? Observons d'où viennent *mise* (n° 24, score 15, fréquence 22) et *mettre* (n° 82, score 7, fréquence 17) ; en cooccurrence non aléatoire avec *pied*, *mise* est attesté dans 3 types de syntagmes :

20) idéaux et contribua en particulier à la **mise** sur pied de la Société des Nations où se reflétaient des Chazelle J., La Diplomatie, 1962, p. 31 (16 attestations)
21) l'*Allemagne a obtenus pour la **mise** sur le pied de **guerre** de ses nombreux contingents. Davout L., Projet de réorg. militaire, 1871, p. 7 (5 exemples)
22) la demande **mise** par l'avocat Labori aux pieds de sa majesté comportait trois solutions Baumont M., Affaire Dreyfus, Archives diplomatiques, 1900, p. 256

Près de *pied*, *mettre* est attesté dans les locutions *mettre sur pied, mettre au pied du mur, sur pied d'égalité, sur le même pied :*

23) en bloc et qu'on **met** sur le même pied de simples règles techniques de bonne gestion et des règles Vedel G., Elém de droit constit., 1949, p. 493
24) la menace de l'invasion pour les **mettre** au pied du **mur** : ou transiger, sinon capituler Lefevre G., La Révolution fr., 1963, p. 566
25) est passé maître dans l'art de **mettre** sur pied de tels rassemblements Meynaud J., Groupes de pression en France, 1958, p. 153
26) l'avantage de **mettre** tous les constituants sur un pied d'**égalité** Niggli P., Phases minér. et pétrogr., t. 1, 1938, p. 26
27) et **mit** d'emblée le nouveau service sur un pied de quasi-perfection Rousseau P., Histoire des transports, 1961, p. 59

En dehors de ces locutions, *mettre* cooccurre avec *pied* dans des types de contextes différents, que l'on peut discriminer par les traits sémantiques /humain/, /animal/, /végétal/, /objet fabriqué/[23] :

i) /humain/

28) avancée, qui **met** littéralement le moteur sous les pieds du **conducteur** Tinard H., Automobile, 1951, p. 336
29) le pied de **position** se **met** sur **pointe** à chaque tour pour Bourgat M., Technique de la danse, 1959, p. 109

[23] Nous verrons au chapitre 3 l'importance des ces traits de dimension.

ii) /animal/
30) en courant il [l'animal sauvage] **met** toujours très régulièrement le pied de derrière à la place que celui de devant vient La Hetraie, La chasse, 1945, p. 158
iii) /végétal/
31) qu'elle apprenne à **mettre** l'engrais au pied de cette **plante**, afin que la pluie l'entraîne Hist. inst. et doctr. pedag., 1949, p. 412
iv) /objet/
32) c'est aux pieds de devant que l'on **met** les **ornements** adoptés. Nosban, Manuel menuis. ébénist., marquet., 1857, p. 16
33) **mettez**-les au pied de la **croix**, couvrez les du sang de Monod H., Sermons, fragments et lettres, 1911, p. 209

*Mettre* est sélectionné également dans le corpus Roman et y figure dans les mêmes locutions que dans le corpus Technique et d'autres types aussi : *se mettre sur pied, mettre qqc aux pieds de qqn, se mettre (à faire qqc)* tandis que *mise* n'est pas retenu (on ne rencontre pas les expressions *mise sur pied, mise sur le pied de guerre*), mais que *mis* l'est, en tant que participe adjectivé ou forme d'un temps composé.

**Les exemples du corpus Roman pour la séquence *pied(s) de (d', du)***

Le premier cooccurrent sélectionné est *grue,* (non sélectionné dans le Technique) uniquement attesté dans la locution *faire le pied de grue,* dans 42 occurrences (*cf*. aussi ex. j) :
a) vous m'avez tout l'air de faire le pied de **grue** Murger, Scènes de la vie de bohème, 1869, p. 258

Les scores importants de *grue* (84), des deuxième et troisième cooccurrents, *lit* (67) et *conculcatrice* (51), marquent un lien sémantique fort avec le lemme *pied(s)* : la locution est attestée comme telle depuis le début du XVII$^{è}$s., après avoir d'abord vécu dans sa forme première *faire de la grue* (début XVI$^{è}$s.), et *pied de lit*, très fréquent, surtout dans le syntagme *au pied du/de son lit* est tellement figé qu'il pourrait être un mot composé avec trait d'union. Cette expression rend compte de 94% des 270 cooccurrences de *lit* et de *pied* (ci-dessous des exemples des autres associations). *Grue* ne cooccurre en fait qu'avec la forme singulier du lemme, *lit* également[24], alors que pour *conculcatrice*, c'est le pluriel qui lui est solidaire : le sens du mot l'explique, "celle qui foule aux pieds", au sens figuré, c'est-à-dire "celle qui méprise"[25].
b) **lit** n'est pas défait ; le **couvre**-pied de soie n'a pas été enlevé Gozlan L., Le notaire de Chantilly, 1836, p. 130
c) la rivière mesurait encore soixante à soixante dix **pieds** de **large**, et son **lit** cinq à six **pieds** Verne J., L'île mystrieuse, 1874, p. 233
d) de remonter chez elle ; elle s'endormit aux pieds de *serge, en travers, sur le **lit, rêvant** Zola, La faute de l'abbé Mouret, 1875, p. 1367
e) Alors, éperdue, n'ayant sous la main que les simulacres de la Révolte, les simulacres de la Bêtise, et les simulacres de l'Idolâtrie, elle (la France) les avait jetés aux pieds de la **Vierge Conculcatrice**, comme l'**Antiquité** renversait aux pieds de Jésus les autels des dieux. Bloy L., La femme pauvre, 1897, p. 195.

*Astérion*, nom de fleur, employé uniquement par Flaubert dans les deux versions de *La tentation de Saint Antoine* de 1849 et 1856, est le 5è cooccurrent sélectionné, avec le score important de 36 qui rend compte de son caractère de hapax dans le corpus Roman : isolés dans le corpus Roman, dans un texte qui est plutôt un conte philosophique, ni le syntagme *couronnes d'astérion*, ni celui de *au pied de mes images* ne se rencontrent dans le Technique.
f) elle crie : oui ! oui ! … au pied de mes images, mes **couronnes** d'**astérion** s'**effeuillent** Flaubert G. La tentation de Saint Antoine, 1856, p. 628

[24] Dans le Technique, sur 6 attestations, on rencontre trois fois *pied du lit,* soit 50%.
[25] Pour la compréhension, nous donnons ici le texte plus développé que la fenêtre de 10 mots avant et après le mot pôle *pied(s)* ; c'est l'occasion de préciser que les programmes (v. 1. 2.) retiennent, avant le calcul du score statistique, les mots qui ont une fréquence de cooccurrence avec le mot pôle supérieure à 1 et non les mots de fréquence supérieure à 1 (car dans ce cas, *conculcatrice* attesté une seule fois aurait été éliminé) : comme il y a deux attestations de *pied(s)* dans l'exemple, *conculcatrice* cooccurre avec l'un et l'autre, et a donc une fréquence de cooccurrence de 2, donc il est retenu pour le calcul du score. Ce type de cas est peu fréquent mais se rencontre ; plus souvent on a répétition du cooccurrent près d'une seule attestation du pôle, comme pour *bastion*, exemple 13.

Les cooccurrents suivants, *mur*, *arbre* et *égalité* sont en relation sémantique avec le singulier du lemme, dans des locutions figées[26] : *au pied du mur* qui peut être attestée au sens propre et au sens figuré (*cf.* les exemples), *au pied d'un arbre/de l'arbre/mon arbre/des arbres*[27], et *sur (un, le) pied d'égalité* :

g) quand je l'ai **mis** au pied du **mur** en lui demandant s'il votera, oui Malraux A., Les Conquérants, 1928, p. 97
h) bien que la **borne** fut au pied d'un **mur**, il s'**adossait** pas Montherlant H. de, Le Songe, 1922, p. 165
i) **asseyez**-vous, monsieur, là, au pied de mon **arbre** … non ? je vous assure Bazin R., Le blé qui lève, 1907, p. 216
j) *des *cigales descendait à *suresnes faire le pied de **grue** devant sa boutique, caché derrière un **arbre** Queneau R., Loin de Rueil, 1944, p. 180
k) qu'un pauvre consentît à m'admettre sur un pied d'**égalité** Larbaud V. A. O. Barbabooth, 1913, p. 105
l) elle vivait avec eux sur un pied d'**égalité**, s'**asseyant** à la même **table** Zola E., Madeleine Ferrat, 1868, p. 123

Avec *diamètre*, le contexte est descriptif, *pied* étant alors unité de mesure (comme c'est le cas avec les cooccurrents comme *hauteur* (n° 22), *large* (n° 26), *bau* (n° 28), *haut* (n° 29), *profondeur* (n° 34), *longueur* (n° 37) ; à ce type d'associations appartiennent également *sol* (pour des expressions comme *à x pieds du sol*) et *mesurer*. On notera que tous les exemples de *mesurer* (*X pieds de longueur, épaisseur, long, longueur, haut, large, largeur, profondeur, envergure, bau, distance*) proviennent des romans de J. Verne *(L'Ile mystérieuse, Les enfants du Capitaine Grant, Les 500 millions de la Begum*, soit 24 exemples, *cf.* exemple c) : ils appartiennent au sous-genre "roman d'aventures", qui nécessite plus de précision pour engendrer l'ancrage référentiel, "faire voir" le décor, que le roman psychologique ou le roman épistolaire par exemple. Et on observe que ni *mesurer* ni *diamètre, distance, épaisseur, longueur, hauteur, profondeur, bau, envergure* n'ont été sélectionnés dans le corpus Technique alors qu'on aurait pu en faire l'hypothèse.

m) déployée, elle a quatre ou cinq pieds de **diamètre**. Hugo V., Les travailleurs de la mer, 1866, p. 372

Avec *escalier, mont, montagne* et *muraille, échelle, rocher*, la cooccurrence avec *pied(s)* vient des locutions du type " au pied de", avec de rares exemples atypiques :

n) le long des maisons **grandes** ouvertes, tourna au pied de la volée d'**escaliers** où le **chien** jaune était Moinot P., Le sable vif, 1963, p. 248
o) jours, à la cave, nous avons au pied du **mur** dégagé un e**scalier** de cinq ou six **marches** Bataille M., L'arbre de Noël, 1967, p. 180
p) se posta devant la porte et l'autre au pied de l'**escalier** Bosco H., Le mas Théotime, 1945, p. 196
q) **dunes**, je creusai dans le **sable**, au pied du **rocher** où elle avait rendu l'âme Benoît P. L'Atlantide, 1919, p. 307
r) je n'étais plus qu'à dix pieds du ruisseau : j'avais gagné les **rochers rouges** About E., Le roi des montagnes, 1857, p; 215
s) descendre de **rocher** en **rocher** au pied de la **falaise** liquide ; il s'avança **pieds nus** Beauvoir S., Les mandarins, 1954, p. 219
t) et ne s'**arrêta** que le soir au pied du **mont** *talbot Verne J., Les enfants du capitaine Grant, 1868, p. 107
u) j'avais été avec ma **grand**-mère, au pied d'une **montagne** honorée par les promenades de Goethe Proust M., La recherche. Côté de Guermantes, t. 1, 1920, p. 256
v) le brave **chien** resta donc au pied de la **muraille**, pendant que son maître et ses Verne J., L'île mystérieuse, 1874, p. 27
w) elle écarta le rideau qui **couvrait** le pied de la **couche** et le repoussa contre la **muraille** Flaubert G., Prem. éducation sentimentale, 1845, p. 34
x) au moyen d'une **échelle** de cinq ou six pieds de **longueur**, qu'on lui tendit de la **petite** Stendhal, L'Abbesse de Castro, 1839, p. 201
y) l'ambition, il lui coûtait de retirer son pied du premier **bâton** de l'**échelle** par laquelle il devait Balzac H. de, Les illusions perdues, 1843, p. 69
z) la *teuse veillait au pied de l'**échelle** Zola E., La faute de l'abbé Mouret, 1875, p. 1435

Si *mont* et *montagne* cooccurrent également avec *pied* dans le corpus technique, on remarque que ces mots y sont plus souvent suivis d'un nom propre, fixant la référence (en géographie par exemple), que dans le corpus roman.

[26] Nous donnons un exemple où *arbre* et *pied* cooccurrent différemment, mais comme pour *lit* et *mur*, la grande majorité des attestations concerne la locution figée : si la locution était retenue comme unité polylexicale, les scores de cooccurrence de *lit, mur, arbre* serait calculés seulement à partir de ces exemples "atypiques" et seraient donc plus forts. Pour *rocher*, la proportion est de 14 exemples de *au pied du rocher* sur 26 exemples.
[27] Le syntagme *pied de/d'un arbre* représente 87% des exemples.

On a vu que dans le corpus technique on rencontrait deux fois *pied de céleri* dans un ouvrage médical : dans le Roman, sur 5 attestations de la cooccurrence, trois concernent également la plante[28], et deux le plat désigné par *salade de céleris* et *céleri rémoulade*, associé dans les deux exemples à d'autres mets populaires, *du pâté de campagne*, *du pied de cochon vinaigrette* et *des pieds de porcs*

a') parut au détour de l'**allée**, **tenant** des pieds de **céleri** dans ses mains pleines de **terre**. Pourrat H., Les Vaillant. Château des septs portes, 1922, p. 87

b') **céleri** rémoulade, du pâté de campagne et du pied de **cochon** vinaigrette à la charcuterie du maréchal Leclerc. Dutourd J., Pluche ou l'amour de l'art, 1967, p. 127

c') leurs **bases**, côtelés sur leurs parcours comme des pieds de **rhubarbe**, **cannelés** comme des **céleris** Huysmans J.-K., En route, 1895, p. 47

d') de mangeailles, dont elle raffolait, telles que pieds de porcs, **salade** de **céleris**, miroton Huysmans J.-K., Les sœurs Vatard, 1879, p. 52

Le dix-neuvième cooccurrent sélectionné est *assis* (score 20, *cf*. aussi *asseoir*, n° 50, score 13, *assise* score 3.6, *assises* score 4.2[29]) qui désigne une attitude :

e') *m *l' *ambert , qui s' était **assis** au pied d' un **arbre** et saignait mélancoliquement . ABO.nn1862

f") .. *andré était **assis** , en pyjama , au pied du **lit**. MART.d09

g') des convalescentes, **assises** sur des **chaises** au pied de leur couches, cousaient, vêtues d'une robe Maupassant G. de, Contes et Nouvelles, 1884, p. 260

Dans le Roman, le test a sélectionné également d'autres termes désignant des attitudes d'acteurs humains : *couché, prosterné, prosterner, agenouiller*, etc. qui ne sont pas relevés dans le corpus technique[30]

h') la fille était **prosternée** aux pieds de son père. Balzac H. de, La cousine Bette, 1846, p. 248

i') il se voyait tout **petit**, **agenouillé** au pied du **lit** maternel -ce **lit** même où maintenant Martin du Gard R., Les Thibault, L'été 1914, 1936, p. 1264

En général, les attestations concernent des acteurs humains, quelquefois animés non humains (un chien peut être décrit assis) : mais on rencontre également des exemples où ces mots sont employés à propos d'inanimés, et cela est typique du corpus littéraire :

j') ville voisine, aristocratique et coquette, tapie au pied de son **château**, **prosternée** devant ses châtelains titrés. Colette, La maison de Claudine, 1922, p. 133

On a noté que les exemples 15 et 16, seuls exemples de *prostrée* dans le corpus technique, ressemblaient au "style" littéraire : de fait, ils émanent d'un texte d'histoire, donc de sciences humaines et pas d'un texte technique comme les ouvrages sur le football ou la menuiserie que nous avons cités à plusieurs reprises. Ces deux exemples décrivent une attitude de prosternation, de la Vierge Marie au pied de la croix du Christ (*cf*. 2. 3. 3) et il se trouve que les exemples de *agenouiller* ou *prosterner* dans le corpus Roman décrivent de telles scènes, que nous étudierons au chapitre 3.

***Les contextes des 3 cooccurrents communs parmi les 20 premiers sélectionnés***

Dans le tableau 1, on constate que sur les vingt premiers lemmes sélectionnés, seuls *croix, égalité* et *biche* sont communs aux deux corpus. L'analyse des contextes montre cependant des différences intéressantes au plan sémantique.

[28] Dans *un pied de céleri*, la relation sémantique est du type partie/tout, car *pied* permet de quantifier l'entité indénombrable désignée par *du céleri* (comme *tête* pour *l'ail* etc.).

[29] La forme *assises* peut être substantif : une affaire politique **occasionnait** une sorte d'**attroupement** au pied du double **escalier** qui mène à la cour d'**assises** Balzac H. de, Splendeurs et misères, 1847, p. 631 ; les lames **décroissantes** laissaient à découvert, au pied de la **petite** *douvre, quelques **assises plates** ou peu Hugo V., Les travailleurs de la mer, 1866, p. 244.

[30] On y trouve un exemple de *corps penché en avant* dans le texte sur le football, le deuxième exemple concernant un véhicule d'attelage (Bourdé, Les travaux publics, t. 2, 1929, p. 12).

**Les associations de *pied* et *égalité***

En fait, dans les deux corpus, *égalité* n'est attesté que dans l'expression *sur (un, le) pied d'égalité*, qui est employée aussi bien dans le roman que dans des textes appartenant à différentes disciplines de sciences humaines ou domaines techniques : architecture, histoire de l'art, droit, philosophie, économie ; diplomatique, anthropologie, textes techniques sur la pêche maritime, le charbon, la gravure. On pourrait penser que cette unité polylexicale est complètement figée et donc la faire coder comme telle par le dictionnaire en amont. Mais dans les deux corpus, on s'aperçoit qu'il n'en est rien, et on trouve des adjectifs antéposés et postposés à *égalité* : *parfaite* et *absolue*, qui renforcent le sens de la locution sont attestés dans les deux corpus :

Technique :

l'origine, est présidé alternativement et sur un pied d'**égalité** absolue, par le général de *gaulle et Vedel G. Manuel élém. de droit constit., 1949, p. 270

ses accès seront toujours libres et ouverts sur un pied de parfaite **égalité** aux navires de **guerre** et de commerce Documents hist. contemporaine t. 2, 1959, p. 295

Roman :

des doctrines radicales et entendaient vivre sur le pied d'une **égalité** absolue Reybaud L., Jérôme Paturot, 1842, p. 293

à la condition de vivre sur le pied d'une parfaite **égalité** dans le ménage Gozlan L., Le notaire de Chantilly, 1836, p. 27

On rencontre dans les deux corpus la variation article indéfini ou défini avant *pied* et l'absence de déterminant, mais seulement dans le Roman l'intercalation d'un adverbe :

le masque rude, la brutalité populacière la **mettaient** presque sur un pied d'**égalité** Zola E., La conquête de Plassans, 1874, p. 1084

Dans le Roman seul on trouve un adjectif qualifiant *égalité* par rapport à un énonciateur[31] :

le sauvage et le monsieur sur le même pied d'**égalité** Larbaud V. A. O. Barbabooth, 1913, p. 240

et, tranquillement, sur un pied d'**égalité** charmante, il lui parla de ce dernier Zola E., Au bonheur des dames, 1883, p. 583

avec lui, contre lui, sur un beau pied d'**égalité** haineuse Paysan C., Les feux de la chandeleur, 1966, p. 154

Seule variante dans le Technique : la présence de l'article défini devant *égalité*, dans un énoncé où ce terme est opposé à *la proportionnalité* :

double représentation doit-il se faire sur le pied de l'**égalité** ou sur le pied de la **proportionnalité** Scelle G., Le fédéralisme européen, 1952, p. 18

Pour cette expression, on constate qu'elle est à la fois plus attestée et plus figée dans le corpus technique (score 56, *égalité* en 2è position), et que le corpus Roman (score 29, n° 7) présente des variations propres qui sont importantes si on les rapporte au volume des corpus et au nombre d'exemples : 23 exemples seulement dans le roman (dont le corpus de travail est de 48000 occurrences) pour 37 exemples dans le corpus Technique, de volume 9000 occurrences.

**Les associations de *pied* et *croix***

Dans le corpus technique, *au pied de la croix* est attesté avec le participe *prostrée*[32] à propos de la Vierge Marie dans un texte d'histoire

(...) en lisant, dans le beau livre de Marcel Bataillon sur Erasme et l' Espagne, ce qu'il écrit du succès que connut le thème de la Vierge **prostrée** et pleurante au pied de la **croix** dans l'Espagne pathétique de la fin du XVe siècle (...) Febvre L., Combats pour l'Histoire, 1952, p. 233

L'examen des contextes nous montre que sur 11 exemples du corpus technique, 10 sont de l'ordre du discours littéraire, car issus de textes de sciences humaines comme l'histoire (L. Febvre atteste deux fois le syntagme), l'histoire de l'art, un ouvrage sur les traditions rurales, ou

[31] C'est une opération de rattachement à un foyer énonciatif, importante pour les textes de fiction, qui s'envisage au plan de la dialectique et de la dialogique, *cf.* le glossaire et les chapitres 4, 5, 6. Les textes techniques manifestent au contraire, par différents procédés linguistiques, une volonté d'évacuer la subjectivité et de décrire le monde "réel".

[32] Dans le tableau 1 où les cooccurrents sont lemmatisés, on trouve *prostré* en 10è position, mais seul le féminin est attesté.

bien un texte très particulier, qui ressortit en fait au discours religieux, les *Sermons, fragments et lettres* de H. Monod, qui fournit 6 exemples[33].

pas un seul péché au monde qui, confessé et pleuré au pied de la **croix**, ne trouve pas son pardon H. Monod, Sermons, fragments et lettres, 1911, p. 255

Resté au pied de la **croix**, il a pu seul percer l'ombre complice qui s'élève de toutes parts pour cacher l'assassinat E. Faure, Histoire de l'art, L'art moderne, 1921, p. 67

paroissiens chantent : "*hosannah *filio *david" au pied de la **croix** hosannière Menon-Lecotte, Au village de France, t. 1, 1954, p. 55

Parallèlement, on constate que dans le seul exemple de texte "technique", l'association entre *pied(s)* et *croix* n'est pas due au syntagme *pied de la croix*[34] :

(…) Jean Loque, orfèvre du clergé, qui renouvelle le trésor de Notre Dame de Paris, sous l'Empire, notamment avec un soleil de 3 pieds de haut et une grande **croix** de vermeil, servant aux processions de fêtes solennelles. Grandjean S., L'orfévrerie du XIXè s. en Europe, 1962, p. 83

alors que le corpus Roman atteste, lui, massivement ce syntagme, dans 87% des exemples.

C'est le critère de "discours" qui conditionne la cooccurrence de *croix* et de *pied* : si le corpus technique était "homogène" et ne comprenait que des textes comme ceux qui ont suscité la sélection des premiers cooccurrents différents que nous avons passé en revue (en 2. 3. 2.), *croix* n'aurait pas été attesté de façon statistiquement valide près de *pied*, à la différence de *bielle, circonvolution, coup, réparation*. L'étude du corpus hétérogène qualifié de "technique" mais comprenant des textes de sciences humaines (dont la langue s'apparente à celle des textes littéraires[35]) et des manuels consacrés à un domaine technique met en évidence que le syntagme *au pied de la croix* est propre aux discours religieux et littéraire, ce que confirme également la présence de cet énoncé de Zola, dans le Roman, tout à fait proche de celui de L. Febvre :

la **Vierge pleurant** au pied de la **croix** Zola, Le Rêve,1888, p. 120.

**Les associations de *pied* et *biche***

Dans le corpus technique, plusieurs énoncés attestent le syntagme *pied de biche* qui désigne différents outils, et que l'on trouve aussi pour qualifier une sorte de pied de meuble :

> *[à l'époque de la Régence] Les tables de milieu, au contraire, perdent leur entrejambe, et les pieds en balustre et en gaine sont de plus en plus rares, car la ligne droite commence à être évincée. On préfère les pieds cambrés, à double inflexion en S allongé, ou les pieds en console dits aussi* pieds de biche. *A l'époque précédente,* le pied de biche *se termine presque obligatoirement par le sabot de l'animal dont il est sensé représenter le membre postérieur ; sous la Régence* le pied de biche *se termine le plus souvent par une volute reposant sur un dé.* Viaux J., Le Meuble en France, 1962 p. 85

> *Les pinces ou leviers sont des barres de fer droites ou recourbées, terminées à une extrémité en pointe, en ciseaux ou en* pied de biche. Bourdé P., Les Travaux Publics, 1928, t. 1, p. 103

Ce syntagme se retrouve dans le corpus littéraire avec deux passages qui réfèrent à un type de sonnette :

> *Un perron branlant de trois marches, une porte à judas de cuivre, avec* sonnette à pied de biche*, et derrière, un corridor sur lequel s'ouvraient, à droite, deux grandes pièces et, à gauche, deux petites.* Huysmans J-K, L'Oblat, 1903, p. 98

[33] Les contextes sont religieux sans ambiguïté : on y trouve *pleurer au pied de la croix, ouvrir son âme, préparez votre conscience, elle est chargée de vos péchés, le sang de Christ nous purifie de tout péché*, etc.

[34] C'est le cas aussi pour la cooccurrence de *pied* et *mur*, attestée 5 fois : les 4 énoncés de sciences humaines ont l'expression figée mais dans le seul texte technique *pied* et *mur* ne sont pas en relation sémantique.

[35] On notera qu'en effet les textes didactiques de la première moitié du XXès. que l'on rencontre dans Frantext sont marqués par la formation littéraire de leurs auteurs, et par le fait que tout niveau de langue "soutenu" se doit de référer au style des "humanités", comme par exemple cet énoncé en géographie : *Rouen se serre au pied de sa falaise*. Vidal de La Blache, 1908, p. 177.

Le dispositif de la sonnette peut figurer de façon plus ou moins réaliste le pied de l'animal, mais dans l'exemple ci-dessous, le narrateur insiste sur le fait qu'il a été conçu avec "un vrai pied de biche" :

> *A la porte de Boulard, il y a* un pied de biche**,** *un vrai* pied de biche, *avec la corne et les poils, que l'on tire, et qui actionne une clochette dans les profondeurs du logis. Cet objet, qu'il a déniché jadis à la Foire à la ferraille et qu'il a posé avec le plus grand soin, a, naturellement un caractère sacré. C'est le plus beau* pied de biche *de l'univers et en même temps un ustensile infiniment plus commode qu'une sonnette électrique.* Dutourd J., Pluche ou l'Amour de l'Art, 1967, p. 142

Dans cet exemple, et malgré l'assertion "un vrai pied de biche", l'analyse sémantique valide dans ce syntagme le trait /objet fabriqué/, récurrent dans *ustensile*, *tirer*, *actionner*, *dénicher*, *objet*, et nonobstant la présence de la corne et des poils -et donc l'impression référentielle[36] qu'ils produisent de "faire voir" un membre de l'animal-, le trait /animal/ ne peut lui être attribué. On peut donc dire que dans le corpus technique et dans quelques énoncés du corpus roman, on a affaire à la lexie *pied(-)de(-)biche*, unité fonctionnelle de communication, attestée comme telle dans les dictionnaires et que le système automatique doit reconnaître.

Mais un autre type de séquence liant *biche* et *pied* se rencontre dans le corpus Roman uniquement et le syntagme *le pied de la biche* s'y insère dans des aires sémantiques particulières :

> *D'un taxi qu'elle arrêta soudain, une jeune femme fit signe à un second taxi, descendit du premier à la hâte, paya sans réclamer sa monnaie, sauta dans le second, et disparut. Nous venions d'assister au relais d'une âme agitée, d'une kleptomane poursuivie, d'une adultère surveillée. C'était* le dernier changement de pied de la biche, *avant qu'elle soit atteinte et verse d'abondantes larmes.* Giraudoux J., Bella, 1926, p. 47

> *Quelque fougueux qu'il soit, aucun cheval ne résiste à son poignet nerveux, à cette main molle en apparence et que rien ne lasse.* Elle a le pied de la biche**,** *un petit pied sec et musculeux, sous une grâce d'enveloppe indescriptible. Elle est d'une force à ne rien craindre dans une lutte ; nul homme ne peut la suivre à cheval.* Balzac H. de, Le Lys dans la Vallée, 1836, p. 230

Dans ces deux attestations *biche* est employé à propos d'une femme : le signifié de ce mot comporte ici les traits sémantiques /humain/ et /féminin/ mais aussi /apeuré/ pour le premier exemple, /endurance/, /gracilité/, pour le suivant. En sémantique on dira que la propagation du trait /humain/ virtualise le trait /animal/ qui est inhérent au contenu de *biche* "en langue". En effet, les sèmes inhérents (comme /animal/ pour *biche*) sont définitoires du type[37], mais l'occurrence n'hérite ces traits du type, par défaut, que si le contexte n'y contredit pas[38]. Cet exemple permet d'illustrer la distinction entre "signification" (du type) et "sens" (de la lexie en contexte), entre description statique et description dynamique. En effet, les traits /humain/ et /animal/ s'excluent mutuellement, et l'opération interprétative ne consiste pas à additionner les traits du type et ceux de l'environnement. "C'est au palier du texte que la conception commune de la compositionnalité laisse apparaître le plus clairement ses lacunes : le global y détermine le local et le *recompose*. C'est pourquoi une phrase et *a fortiori* un mot peuvent changer de sens quand se modifie le contexte immédiat et lointain"[39]. Les limites de la définition lexicographique sont également mises en évidence : dans les dictionnaires en général et même dans le TLF qui s'appuie sur le fonds dont est issue la banque Frantext, on ne trouve pas de

[36] V. 1.2.
[37] Pour *biche*, il correspond (grossièrement) au sens 1 des dictionnaires "femelle du cerf". On observera que l'article défini pour *la biche* est employé pour renvoyer aux caractéristiques génériques, celles de "la classe" des biches ; pour l'emploi du défini singulier avant *pied*, v. chapitre 3.
[38] V. le glossaire pour *propagation, sème inhérent, type* et *occurrence, virtualiser*.
[39] Rastier F. dans Rastier F. et *alii*, 1994, p. 172.

subdivision qui rende compte de l'ensemble de traits /humain/ /féminin/ /gracile/ /endurant/ ou /humain/ /féminin/ /apeuré/[40]. Remarquons que ni dans le corpus Technique, ni dans le corpus Roman, *biche* ne réalise les traits génériques /animal/ /animé/, traits qu'un étiquetage sémantique automatique lui aurait pourtant affecté. La tentation serait grande, de même, de donner à *pied* le trait /animé/ ou /partie du corps/. On objectera que le problème de fait est celui de la reconnaissance des unités polylexicales qui doivent se trouver dans le dictionnaire de formes auquel sont liés les traitements statistiques et informatiques : mais si *pied de biche* peut effectivement être reconnu à priori (le trait d'union est cependant facultatif), le problème des figements est plus complexe car lié au genre de textes, comme nous en avons eu un exemple avec la locution *sur pied d'égalité* : la séquence "le pied de la biche" ne pouvant être considérée comme figée, le trait /animal/ aurait été affecté à *biche* par un traitement automatique. La reconnaissance des "unités de langue" d'un corpus passe par la prise en compte des critères de discours et de genres.

## 2. 2. LA VARIABLE 'CHAMP GENERIQUE'

### 2. 2. 1. Les réseaux associatifs de *amour* dans le corpus Roman et dans le corpus Poésie

Si les associations varient entre deux discours, peut-on de même repérer des différences entre champs génériques ? Nous employons les distinctions de *discours, champ générique, genre et sous-genre* proposées par Malrieu D. et Rastier F. (à paraître dans T. A. L.) dans leurs travaux proposant une typologie des textes fondée sur des caractéristiques homogènes et linguistiques[41] :"Nous distinguons quatre niveaux hiérarchiques supérieurs au texte : les *discours* (ex. juridique *vs* littéraire *vs* essayiste *vs* scientifique), les *champs génériques* (ex. théâtre, poésie, genres narratifs), les *genres* proprement dits (ex. comédie, roman "sérieux", roman policier, nouvelles, contes, mémoires et récits de voyage), les *sous-genres* (ex. roman par lettres)."

Ayant observé des différences notables dans les textes selon le type de discours, littéraire *vs* non-littéraire[42], nous avons voulu tester le niveau hiérarchique inférieur au discours : le test de l'écart réduit est appliqué à un corpus de travail constitué des séquences de 10 mots avant et après le mot *amour*, au singulier, dans deux champs génériques, le roman et la poésie, pour vérifier si les associations entre cooccurrent et mot pôle sont comparables.

Le corpus Poésie de Frantext est peu important quantitativement (4,6 millions de signes) et poserait des problèmes de représentativité pour des études de genres (comme la ballade) ou d'auteurs. De plus, le fait que l'ensemble textuel ne soit pas enrichi rend impossible un

[40] La partie B2 de l'article du TLF se distingue en fait par les traits /humain/ et /féminin/ de la partie A "femelle du cerf" et B1 "technologie" (qui enregistre *table à pieds de biche, pied-de-biche, couleur pied de biche*). Cette partie atteste les syntagmes *avoir un cou, des yeux de biche, vive comme une biche*, ainsi que l'hypocoristique *ma biche* et le sens argotique vieilli de "femme entretenue" : ce sont bien les traits de dimension qui fondent le plan de l'article mais pas de façon explicite parce que le TLF ne s'appuie pas sur une théorie sémantique. On notera que le trait /peur/ n'est attesté que dans le syntagme *des yeux de biche, apeurés et tendres*.

[41] Cet article porte sur la caractérisation morphosyntaxique d'un corpus de 2500 textes complets, classé par genres et discours et étiqueté par 251 types d'étiquettes.

[42] Ces différences auraient été plus sensibles encore si le corpus technique de Frantext avait été plus homogène, comme signalé ; on a vu qu'un texte du discours religieux se signalait par des associations bien différentes des autres textes.

traitement automatique puisqu'on n'a même pas accès au titre ni aux limites des pièces dans l'édition, ni à leur caractérisation minimale pour délimiter un corpus compatible avec sa recherche, sans parler des autre types de balises internes au texte (comme celles préconisées par la TEI[43] que l'on peut adapter à son propos).

L'expérience de confrontation entre le "discours littéraire" et le "discours technique" nous a permis de souligner certaines des exigences philologiques auxquelles doit se conformer une banque de textes comme Frantext et de mettre en évidence des lacunes criantes qui n'empêchent cependant pas les trouvailles dues au changement d'échelle que représente l'accès à des corpus vastes. Cette recherche contrastive entre roman et poésie, autour d'un mot pôle, permet néanmoins de montrer qu'on n'a pas la même probabilité de trouver telle association dans un champ générique et dans l'autre. Les résultats du test sont représentés dans les deux graphiques ci-dessous : la différence de volume des corpus se manifeste dans la taille des éléments des diagrammes[44] où la même échelle de représentation a été utilisée et où les bâtonnets figurent les scores statistiques.

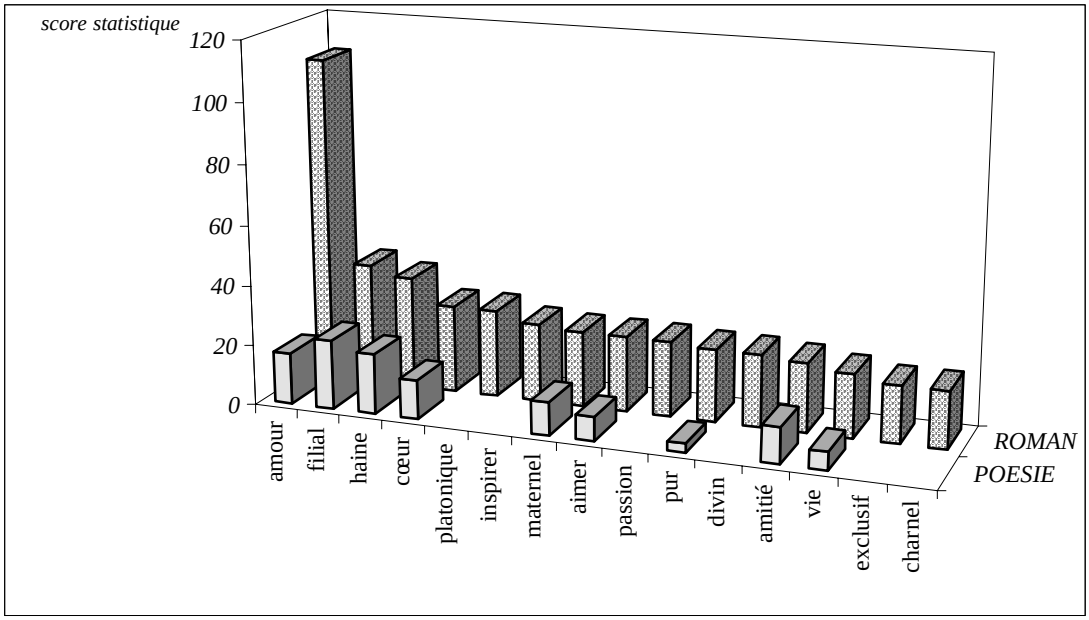


*Diagramme 1 : les scores des 15 premiers cooccurrents sélectionnés près de* amour *dans le Roman et leur score en Poésie*

[43] Le sous-genre devrait apparaître dans un *en-tête* (cf. TEI), ainsi que les précisions quant aux éditions du vivant de l'auteur, etc. : ce sont des variables globales qui déterminent en partie le plan local, *cf.* Malrieu D. et Rastier F., *loc. cit.* L'usager doit pouvoir reconnaître et isoler chaque vers (est-il complet ou segmenté ? quel est son numéro d'ordre ?), les groupes de vers (1er quatrain), pouvoir étudier les types de vers (vers simple ou composé, vers libre, vers libéré), les mots à la césure, en fin de vers, etc. v. Sperberg-McQueen C. M., Burnard L, 1999. Dans Frantext la présentation des pièces poétiques est très lacunaire ; dans toutes les banques, l'usager devrait avoir accès au mode image en même temps qu'au mode texte (sur un texte enrichi). Pour tous les types de textes, ce serait important mais encore plus en poésie : que l'on songe aux *Calligrammes* d'Apollinaire.

[44] Elle explique que le corpus Poésie se trouve toujours figuré en premier.

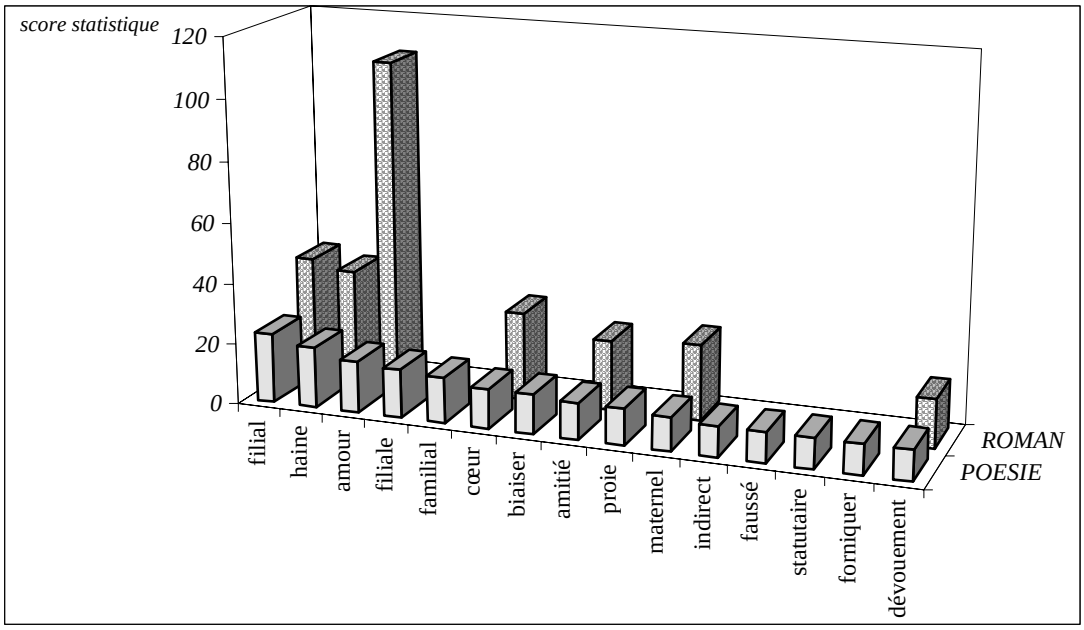


*Diagramme 2 : les scores des 15 premiers cooccurrents sélectionnés près de* amour *en poésie et leur score dans le Roman*

Sur les quinze premiers cooccurrents sélectionnés, il y a recouvrement pour neuf des premiers du corpus Roman, et pour sept du corpus Poésie. Parmi les premiers sélectionnés du corpus Roman, *platonique, inspirer, passion, divin, exclusif* et *charnel* ne sont pas sélectionnés en tête dans le corpus Poésie, et parmi ceux-ci, comme le montre le tableau comparatif en annexe, seul *divin* est cooccurrent de *amour* (avec le score 7) dans le corpus Poésie.

### 2. 2. 2. Syntagmes et relations sémantiques

#### *Les cooccurrents différents*

##### Les cooccurrences dans le corpus Roman

On remarque que les syntagmes *amour platonique, inspirer (de l'amour/un amour …) amour exclusif* appartiennent à la langue du roman, comme, avec des scores inférieurs pour la cooccurrence des éléments *déclaration d'amour, amour éternel, mariage d'amour, ressentir de l'amour/un amour …, amour conjugal* ou *charnel, véritable, naissant, les délices de l'amour, preuve d'amour*[45], etc.…) : or nous n'avons pas conscience en disant ou lisant ces syntagmes qu'ils soient "spécialisés" et ces recherches contrastives permettent de montrer que "la langue une" est une abstraction des linguistes qui néglige d'une part la dimension du textes, et d'autre part le fait que les textes sont des objets culturels, inclus dans des pratiques sociales[46]. Au plan des sentiments associés à l'amour, si *haine, amitié, aimer, tendresse, cœur* sont relevés avec des

[45] Comme nous l'avons signalé dans la partie 2. 3., la cooccurrence peut se trouver dans une autre forme que le syntagme ci-dessus, mais celui-ci rend compte de la majorité des attestations dans ce champ générique où les figements sont importants ; le score du syntagme serait beaucoup plus important (pratiquement 10 fois plus, pour une illustration, v. chapitre 6). Par exemple à côté de *mariage d'amour* on rencontre des énoncés comme : *le mariage ne saurait accroître notre amour* et *le mariage produit l'amour comme un pêcher une pêche.*

[46] Il y aurait lieu de revoir la distinction "langue générale" *vs* "langues de spécialité" ? Il y a plus de normes sociales implicites à l'œuvre dans toutes les sortes de discours que nous ne le pensons : quand nous aurons à notre disposition des corpus électroniques suffisamment documentés, nous pourrons suivre la diffusion de mots, de syntagmes (et d'ensembles de traits sémantiques, *cf.* chapitres 3, 4, 5) dans des types de textes différents de leur origine (par exemple de quoi est faite "la langue" du journalisme, de la publicité ?).

scores plus ou moins importants près de *amour* dans les deux corpus, on remarque que *passion, passionné et jalousie, ambition, chagrin, orgueil, oubli, renoncement, révulsion, vanité, égoïsme* sont statistiquement significatifs seulement dans le roman[47]. Il ne faut pas en conclure rapidement que "la chose à dire" est obligatoirement différente mais cela signifie que le mode d'expression, de lexicalisation diffère : par exemple, en poésie on n'associe pas *jalousie* à *amour*, mais *jaloux*, épithète de *amour* se rencontre (comme d'ailleurs dans le roman).

Dans le tableau en annexe, on verra que figure en 11è position l'adjectif *irrétractable*, que nous avons supprimé du diagramme des quinze premiers cooccurrents du corpus Roman, pour faire le parallèle avec celui du corpus Poésie, où les cooccurrents hapax n'apparaissent qu'à partir du rang 15 : en effet, *irrétractable* n'est attesté qu'une fois dans le corpus de référence mais a un score de cooccurrence de 2 parce qu'il figure dans un énoncé ou le mot pôle est répété[48].

**Les cooccurrences dans le corpus Poésie**

*Amour* est un nom qui a le double genre et on dit généralement qu'il n'est féminin qu'en poésie : cela se vérifie dans les corpus de travail. Dans le corpus Poésie, le genre féminin de *amour* est à l'origine de la sélection de *filiale* à côté de *filial*[49], forme commune aux deux corpus ; les autres adjectifs employés avec les deux genres du mot pôle sont *divin, sensuel, éternel* ainsi que *maternel* et *fraternel* (chez Péguy). *Familial* n'est en cooccurrence avec *amour*, qu'il qualifie, que dans un énoncé où le mot pôle est répété trois fois, chez Péguy, la répétition servant à rendre le mouvement litanique : le score de cooccurrence en est doublé, comme nous l'avons déjà signalé[50].

Parmi les autres cooccurrents particuliers au corpus Poésie, *biaiser* est dû à deux pièces de Verlaine où il cooccurre deux fois avec le mot pôle[51], et on rencontre *amour indirect* chez Michelet[52] à propos de l'insecte ; *faussé* et *statutaire*, qualifiant *amour*, sont des associations de Péguy[53], comme *déharnachement*[54] qui figure dans trois strophes de *Eve*.

---

[47] Sur *amour* et *ambition* dans le roman, v. Ehrlich D., 1995.

[48] Les citations sont telles que dans Frantext. *Toi qui me fait tant regretter d'avoir écrit cette phrase absurde et irrétractable sur l'amour, le seul amour, "tel qu'il ne peut être qu'à toute épreuve"*. Breton A., Nadja, 1928 p. 149.

[49] Comme nous l'avons signalé, les traitements, réalisés ici avec lemmatisation des cooccurrents ne rattachent pas *filiale* au lemme *filial*, à cause de l'existence du substantif féminin. Au plan de la forme on rencontre aussi *dévoûment* dans le corpus Poésie, qui n'est pas attesté dans le Roman.

[50] *C'est le mouvement propre, le mouvement naturel de notre amour. De notre amour humain, de notre amour familial, de notre amour filial. Il y a l'église triomphante. Nous devons tâcher d'en être. Il n'y a pas à s'en cacher. Il n'y a pas à faire le modeste*. Péguy Ch., Mystère de la charité de J. d'Arc, 1910, p. 67-68.

[51] *Il faut oser m'aimer ! oui, mon amour monte sans biaiser jusqu'où ne grimpe pas ton pauvre amour de chèvre, et t'emportera comme un aigle vole un lièvre, vers des serpolets qu'un ciel cher vient arroser !* Verlaine P. Œuvres poétiques complètes. Sagesse II, IV, 1896, p. 268-69 *; Absolution sainte savourée avec crainte d'en être indigne encor, d'en peut-être abuser. Rentrée emmi le monde et son horreur profonde avec un cœur d'amour qui ne sait biaiser, car c'est l'amour divine qui prévoit et devine les pièges, le manège et les tours du *péché.* Id. Liturgies intimes. XX complies en ville, p. 753-55.

[52] *Ce que le vent fait au hasard, jetant, par ondées, par caprice, les éléments générateurs, l'insecte le fait par amour, amour direct de son espèce, amour indirect et confus de cette aimable auxiliaire qui l'accueille et qui le nourrit, qui nourrira même encore ses œufs après lui et continuera sa maternité.* Michelet J., L'insecte, 1857, Livre 3, Soc. des insectes p. 317. Ce texte en prose, atypique, descriptif, poétique et philosophique à la fois, serait plutôt à classer comme "essai".

[53] *L'amour le plus dur et le plus statutaire* rime avec *l'amour le plus pur et le plus salut*aire dans la Prière de déférence de La Tapisserie de Notre-Dame, de Ch. Péguy, 1913, p. 703-704 : *Comment peux-tu les aimer. D'une amour mentie, d' une amour trahie et qui se trahit soi-même, qui se trahit perpétuellement soi-même, d'une amour faussée. Toute droiture est gauchie à présent. Tu mens par le son de ta voix. Tu mens par le regard de tes yeux. Tout s'est à jamais faussé dans ton âme. Et tout s'est à jamais faussé dans ta vie : faussée l'amour filiale et faussée l'amour fraternelle. Id.*, Mystère de la Charité de J. d'Arc, 1910, p. 60.

Dans le tableau des cinquante premiers on peut s'interroger sur *mourre*, qui vient en 16è position : il résulte d'un jeu de mot phonique sur *amour* et cet exemple atteste un autre cooccurrent inattendu, *nombril*, relevé uniquement dans ce texte[55].

*Mal*, cooccurrent n°34, est un ancien adjectif, de même sens que *mauvais* (et supplanté par lui), qui figure 4 fois dans la même pièce chez Moréas, dans le vers-refrain "amour occit mon cœur de male lance"[56]. Cette pièce, une "cantilène" d'inspiration médiévale, a été écrite à l'époque où Moréas, après avoir appartenu au mouvement symboliste, quittait ses amis pour fonder l'école romane, préconisant le retour à la tradition française, avec des formes fixes et des thèmes du Moyen Age et du XVIès. : on y trouve des mots d'ancien et moyen français comme *bracet, samit, rubacelle, escarcelle, chapel, parement, destrier, caparaçonnement, précellence, nonchaloir, allègement.*

*Allégresse, hymen, hyménée*[57]*, igné, incendier, rets, rossignol, soupir* sont des cooccurrents de *amour* dans le corpus poésie uniquement ; de même on y rencontre les syntagmes *dévoré d'amour et cœur dévoré/brûlé/consumé d'amour, la proie de l'amour*[58]*/en proie à l'amour* (avec qualificatif), à côté de *l'amour des proies fraîches*, syntagme attesté 28 fois, à propos de fauves, chez H. de Montherlant[59].

***Les cooccurrents communs aux deux corpus***

Dans le tableau des 50 premiers cooccurrents (en annexe), on voit que la proportion de cooccurrents communs aux deux corpus est plus importante que ce que nous avions constaté pour les corpus Technique et Roman : 24 et 25 cooccurrents communs contre 16 et 13. C'est un effet de l'homogénéité de discours puisque nous avons affaire dans les deux cas à des corpus de discours littéraire. Cela va dans le sens des conclusions de Malrieu D. et Rastier F. (*loc. cit*.) dont les recherches (à partir de variables morphosyntaxiques) ont bien montré que "les distances entre les discours sont plus élevées qu'entre les champs génériques, lesquelles sont à leur tour plus grandes qu'entre les genres" et que "les variables discriminantes à un niveau ne le sont pas nécessairement aux autres", ce qui justifie la hiérarchie en quatre niveaux proposée.

Comme nous l'avons noté dans d'autres expériences sur le roman, un des premiers mots sélectionnés par le test statistique est le mot pris comme pivot de recherche, qui a, statistiquement, une probabilité pourtant plus faible de se trouver en position de cooccurrent : mais dans le discours littéraire, la répétition est un procédé à fonctions multiples, même si elle est "fondamentalement anti-narrative"[60], nous avons observé qu'elle sert tout particulièrement à noter le trait /intensité/[61] dans le roman, ce qui peut être le cas en poésie, où elle peut provenir d'autre part de contraintes prosodiques (rime, assonance, rythme).

---

[54] *Le déharnachement de tendresse et d'amour sur le parvis commun posait une humble trêve (…) Le déharnachement de rudesse et d'amour sur le double parvis posait une heure brève (…) Le déharnachement d'allégresse et d'amour sur le double parvis posait une heure brève*. *Id*. Les Tapisseries, Eve, 1913, pp. 820-821.

[55] *Les humains savent tant de jeux l' amour la mourre l' amour jeu des nombrils ou jeu de la grande oie la mourre jeu du nombre illusoire des doigts seigneur faites seigneur qu'un jour je m'énamoure j'attends celle qui me tendra ses doigts menus.* Apollinaire G., Alcools, 1913, pp. 98-101 ; *Le soleil en dansant remuait son nombril et soudain le printemps d'amour et d'héroïsme amena par la main un jeune jour d'avril. Id., ibid*., pp. 86-89.

[56] Moréas J., Les cantilènes. Tidogolain, Premières poésies, 1886, p. 218.

[57] Ils signifient "mariage" mais on ne les rencontre pas dans les syntagmes **hymen(ée) d'amour* qui seraient parallèles à *mariage d'amour* du corpus roman.

[58] *Dans le domaine irréconciliable de la surréalité, l'homme privilégié ne pouvant être que la proie gracieuse de sa dévorante raison de vivre : l'amour*… Char R., Marteau sans maître. Moulin, 1945, Artine 1930, p. 23.

[59] Dans le recueil intitulé : Encore un instant de bonheur (1934, pp. 700-701).

[60] Molinié G.,1986, p. 164 ; on notera que ce procédé est bien plus utilisé dans le genre romanesque qu'en poésie, la recherche restant à faire pour d'autres genres.

[61] Il faut alors l'étudier en fonction de son origine, le foyer énonciatif auquel elle se rapporte.

Mais *amour*, si l'on en juge par un dictionnaire d'usage courant, peut se trouver dans au moins 8 "sens" différents, et en effet, nous constatons des contextes différents, selon qu'il s'agit, dans les deux champs génériques, du sentiment entre parents et progéniture (sélection de *filial, familial, maternel, maternité, paternel, foyer, famille*), de l'amour homme-femme (*femme, fille, amant, physique, sensuel, sexuel, chaste, caresse, forniquer, acte d'amour*, etc.), de l'amour de Dieu (*divin*), du dieu amour[62], du terme d'adresse à l'être aimé (*cher cœur ! cher amour !*), etc[63]. De même, *haine, aimer, amitié, vie*, cooccurrents sélectionnés dans les deux corpus figurent dans toutes sortes de contextes qu'on ne peut étudier en détail ici.

Au plan de la ponctuation, on remarque que dans les deux champs, le point d'interrogation a été sélectionné avec des scores importants, 13 et 7, rendant compte de la présence du mot dans des contextes marqués par l'énonciation représentée et la subjectivité[64] ; ce critère explique également la sélection de la virgule (de score plus important en poésie cependant) qui réfère à une prosodie brisée dans les deux champs.

Mais même pour les cooccurrents communs, des différences d'emploi s'observent : on dit *consumé d'amour* en poésie, *consumé par l'amour, se consumer d'amour* dans le Roman. Une piste de recherche serait la place de l'adjectif selon le champ : parmi les adjectifs sélectionnés comme cooccurrents, nombreux sont ceux qui sont plutôt antéposés en poésie (*brûlant amour*) et postposés dans le Roman (*amour brûlant*). Une autre direction serait de vérifier si la courbe du poids respectif des cooccurrents suit la courbe intonationnelle, comme le laisse suggérer la représentation de trois énoncés, du corpus Roman et du corpus Poésie, que l'on peut voir en annexe du chapitre ; cela serait possible grâce aux outils de transcription graphème-phonème[65].

****

Ce type d'expérience, en sémantique assistée de corpus, montre l'importance des normes de discours et de champ générique : un même "contenu sémantique" peut être associé à une notion donnée (comme l'amour) mais les mots pour lexicaliser ce contenu et les procédés linguistiques et stylistiques peuvent varier notablement d'un discours, d'un champ générique, d'un genre à un autre, dans des pratiques situées et toujours particulières. Dans les recherches sur *pied*, comme dans celles sur *amour*, on a vu que des éléments interprétables seulement au palier inférieur, (champ générique, genre et sous-genres) étaient dégagés par le test : thématique fixe et particularisée dans les textes sur le football ou la médecine, normes de sous-genre à propos de l'ancrage référentiel dans le roman d'aventures, normes poétiques et stylistiques chez Moréas, Péguy, Michelet, Montherlant dont les associations avec le mot pôle sont uniques. On a pu observer comment le contexte remanie les signifiés au plan local, ce qui relativise la possibilité de codage sémantique de corpus à partir des dictionnaires ou de réseaux comme WordNet, et qui met en évidence les limites de la linguistique du signe.

---

[62] *Ou bien c'est, ouvrant à deux mains le compas, Uranie, à la ressemblance de Vénus, quand elle enseigne, lui bandant son arc, l'amour.* Claudel P., Cinq grandes odes, 1910, pp. 222-223.

[63] Cependant le sens "goût très vif pour quelque chose" appartient pratiquement au Roman (on y trouve *l'amour de l'art, du beau, de la musique, de la science, de la gloire*, etc.) : en poésie il ne se rencontre que dans le syntagme *l'amour des proies fraîches*, chez Montherlant, où il ne s'agit pas d'acteur ayant le trait /humain/.

[64] L'article cité de Malrieu D. et Rastier F. met également en évidence que "la représentation de l'interlocution est partout discriminante de façon cruciale : elle singularise la littérature par rapport aux autres discours et sépare le théâtre de la poésie et du roman".

[65] V. Yvon F. et *alii*, 1998 et Beaudoin V. (à paraître).

Un dictionnaire courant et synthétique comme le *Petit Robert* affiche sur la première de couverture qu'il traite de 60 000 mots et de leurs 300 000 sens. L'identité à soi du mot n'existant pas, il paraît préférable d'accepter l'incidence du "principe sémiotique", qui pourrait s'énoncer ainsi : "tout signe plongé dans un discours, un genre, un texte, est susceptible de recevoir de ce "milieu" des transformations imprévisibles de la signification qui lui est attribuée en tant que "type" : toute occurrence doit, de ce fait, être analysée en fonction du contexte (pris au sens large), de ce milieu écologique qui est un système signifiant".

Si ces variables de genres textuels sont importantes pour les corpus électroniques dont la demande sociale est forte[66], une entreprise typologique des textes scientifiques et techniques est indispensable pour coder ces textes de manière à pouvoir les utiliser en format électronique et à trouver du "sens supplémentaire" de leur réunion en corpus. Et pour tous les types de textes, la question du balisage est cruciale et elle commence par un bon repérage dans une banque de données bibliographiques permettant à l'usager d'avoir accès à toutes sortes d'éléments indispensables pour constituer un corpus pertinent pour sa recherche. "La typologie des genres textuels paraît indispensable pour les traitements automatiques. Soit en général, car l'analyse des corpus en situation montre que le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres. Les systèmes d'analyse et de génération doivent tenir compte de ces spécificités. Les projets de systèmes universels sont ainsi irréalistes, linguistiquement parlant (*en note* : ils reposent en effet sur le préjugé que la langue est homogène et identique à elle-même dans tous les textes et dans toutes les situations de communication). Soit en particulier, car les genres sont déterminés par des pratiques sociales spécifiques, dans lesquelles les applications informatiques prennent place. Elles doivent donc tenir compte de ces contraintes propres à ces pratiques où elles s'insèrent"[67].

## 2. 3. RESULTATS, PROBLEMES, PROPOSITIONS

1) L'importance des normes de discours et de genre incite à reprendre la question de la discrétisation des unités fonctionnelles dans ce cadre :

- on peut approcher ce problème en combinant des ressources existantes (dictionnaires informatisés de référence, lexiques mis au point sur d'autres corpus, des outils élaborés pour l'aide à la lexicographie comme Sextant, etc.), avec l'observation des scores statistiques des voisins de chaque forme graphique et en comparant les fichiers, d'abord de manière automatique, puis en affinant "à la main".
- Il est probable que des seuils de scores qui ont trait au figement puissent être déterminés, si on opère sur des corpus homogènes ; pour des textes techniques, le niveau supérieur sera le domaine, et les scores les plus importants concerneront en grande partie les "mots du domaine"

[66] La poétique généralisée, étudiant tous les types de textes, et pas seulement les littéraires, que préconise F. Rastier "commence au palier des discours. (…) Chaque groupe de pratiques sociales correspondant à un discours se divise en activités spécifiques (…) qui ont chacune leurs genres. Par exemple dans le discours médical, on peut distinguer les genres écrits dont dispose un professeur des hôpitaux dans sa pratique professionnelle : ils sont au nombre de trois, le résumé d'observation clinique, l'article scientifique et la lettre au collègue". Rastier F., Pincemin B., 1999, p. 99.

[67] Rastier F. dans Rastier F. et *alii*, 1994, p. 176.

(comme *ballon, but, surface, réparation* sortaient pour le texte sur le football, et *carotte, veau* pour la cuisine).

- En accumulant des "lexiques vides" (pour reprendre la formule de Sinclair), constitués de cette façon dans des corpus homogènes, on pourra obtenir à terme des répertoires d'unités fonctionnelles assortis de la mention de leur origine, de manière à conserver ces acquis pour d'autres recherches : certaines unités seront propres à un très petit nombre de textes, voire un seul, et d'autres (la "langue générale") se rencontreront partout ou presque, avec des variantes pour certains traits (nombre, genre, temps verbaux, modes, traits de dimension comme /humain/ *vs* /animal/). Un étiquetage multiple sera affecté à ces unités (partie du discours, transcription phonologique, discours, genre, domaine, appartenance à des collocations, types d'expansion des collocations, etc.), et des traits sémantiques de généralité supérieure (discours, domaine, dimension, taxème éventuellement), ce qui permet d'éviter des ambiguïtés en discriminant grossièrement les signifiés.
- Des règles d'héritage de traits pourront être constituées progressivement, testées et améliorées par apprentissage dans des corpus *ad hoc*.

2) Les recommandations de la philologie numérique doivent servir à améliorer l'état des textes numérisés et des banques textuelles, à commencer par Frantext dont on a vu les nombreuses lacunes :

- documenter les textes, pour que l'usager puisse constituer des corpus adaptés à ses recherches ; préciser dans l'en-tête TEI si la banque propose plusieurs éditions du même texte, quelle édition est celle qui fait foi, au plan philologique
- coder la structure "matérielle" des textes (subdivisions internes, pages, paragraphes, didascalies et textes de chaque personnage pour le théâtre, titre des poèmes d'un recueil, etc.), distinguer les mots étrangers, le péri-texte, le para-texte, l'apparat critique, etc.
- généraliser la possibilité d'avoir accès au texte à la fois en mode page et en mode image
- donner des outils de pondération statistique à l'usager, en lui permettant de définir son corpus de référence comme son corpus de travail
- dans le cadre de stations de travail, accès à des ouvrages encyclopédiques, d'histoire littéraire, des dictionnaires d'état de langue ancien et de langues étrangères.